Agricultura Predictiva

Modelos de Machine Learning en Agricultura

Predicción de Rendimiento y Detección de Enfermedades

Juan Jose Van de venter M. - jjvanv @eafit.edu.co

Luis Felipe Arciniegas Espitia - lfarciniee@eafit.edu.co

Departamento de

Ingeniería Agronómica

Universidad EAFIT, Medellín

Antioquia, Colombia

Nov de 2025

Resumen

El aprendizaje automático (Machine Learning, ML) se ha consolidado como una herramienta clave en la agricultura inteligente por su capacidad para modelar relaciones complejas entre variables climáticas, de manejo y rendimiento. Este estudio evaluó y comparó distintos modelos de ML aplicados a dos escenarios: **predicción del rendimiento agrícola** y **detección de enfermedades en aguacate Hass**.

En la primera parte, se utilizaron datos reales del *Crop Yield Prediction Dataset* (Kaggle–FAO), aplicando limpieza, imputación jerárquica y normalización. Se entrenaron tres modelos de regresión: **Regresión Lineal**, **Random Forest** y **Gradient Boosting**, con división 80/20 y métricas MAE, RMSE y R^2. El modelo **Random Forest** mostró el mejor desempeño ($R^2 = 0.869$), confirmando la superioridad de los métodos de ensamble frente a los lineales.

En la segunda parte, se generó un **dataset sintético** con variables climáticas simuladas (1990–2010) para clasificar la presencia de enfermedad según precipitación mensual ≥ 700 mm. Los modelos **SVM-RBF**, **KNN** y **Random Forest** fueron evaluados mediante precisión y recall; el **SVM-RBF** resultó más eficaz para reducir falsos negativos.

Los resultados evidencian que los modelos de ML son herramientas robustas para **anticipar rendimientos y riesgos fitosanitarios**, optimizando la toma de decisiones en agricultura sostenible.

Introducción

La agricultura moderna enfrenta el desafío de mantener su productividad bajo condiciones de cambio climático cada vez más extremas, optimizando simultáneamente el uso de recursos como agua, fertilizantes y pesticidas. En este contexto, la **agricultura inteligente**, basada en tecnologías digitales, sensores e inteligencia artificial, se ha consolidado como una estrategia clave para mejorar la eficiencia y sostenibilidad de los sistemas agroproductivos (Liakos et al., 2018).

Dentro de estas tecnologías, el **aprendizaje automático (Machine Learning, ML)** destaca por su capacidad para identificar patrones complejos y modelar relaciones multivariables entre variables climáticas, de manejo y rendimiento (Sarker, 2021). Los métodos tradicionales de análisis agrícola suelen fallar al asumir relaciones lineales y de independencia, mientras que los modelos de ML permiten capturar interacciones no lineales y jerárquicas entre factores como temperatura, precipitación y uso de insumos (Nayak et al., 2022).

Este estudio tuvo como objetivo **evaluar y comparar modelos de ML aplicados a la agricultura**, estructurando el trabajo en dos partes complementarias. La primera analiza **datos reales** del *Crop Yield Prediction Dataset* (Kaggle–FAO) para predecir rendimientos agrícolas mediante Regresión Lineal, Random Forest y Gradient Boosting. La segunda emplea **datos sintéticos** simulados entre 1990 y 2010 para **detectar enfermedades en aguacate Hass** utilizando KNN, SVM-RBF y Random Forest. En conjunto, ambas fases buscan demostrar cómo el ML puede **anticipar rendimientos y riesgos fitosanitarios**, fortaleciendo la toma de decisiones y promoviendo una agricultura más resiliente y sostenible.

Metodología

Parte 1. Predicción del rendimiento agrícola con datos reales

La base de datos utilizada corresponde al conjunto *"Crop Yield Prediction Dataset"* (Patel, 2020), disponible en Kaggle, el cual reúne información histórica sobre rendimiento de cultivos agrícolas en distintos países y años, junto con variables como temperatura promedio, precipitación anual y uso de pesticidas. Los registros provienen de la FAO y fuentes climáticas oficiales, alcanzando más de 50.000 observaciones y ocho variables de interés.

El procesamiento incluyó limpieza estructural, eliminación de campos redundantes y imputación jerárquica de valores faltantes mediante medianas agrupadas por país y cultivo (Allison, 2001). Posteriormente, se aplicó normalización z-score para las variables numéricas y codificación One-Hot Encoding (OHE) para las categóricas, consolidando un conjunto homogéneo y listo para modelar (FAO, 2022).

Los datos se dividieron en proporción 80/20 (entrenamiento/prueba) y se implementaron tres modelos de aprendizaje supervisado:

1. Regresión Lineal, usada como modelo base por su interpretabilidad (James et al., 2021).
2. Random Forest Regressor, modelo de ensamble paralelo que reduce varianza y captura interacciones no lineales (Breiman, 2001; Belgiu & Drăguţ, 2016).
3. Gradient Boosting Regressor, método de ensamble secuencial que corrige errores residuales, logrando alta precisión en datos tabulares (Friedman, 2001; Nayak et al., 2022).

El desempeño se evaluó mediante tres métricas estándar: Error Absoluto Medio (MAE), Raíz del Error Cuadrático Medio (RMSE) y Coeficiente de Determinación (R^2), que permiten comparar precisión, estabilidad y capacidad explicativa de cada modelo.

Parte 2. Detección de enfermedades con datos sintéticos

Para la segunda fase se generó un dataset sintético de 10.000 registros utilizando la función make_classification, con seis variables (temperatura mínima y máxima, precipitación, humedad, radiación solar y velocidad del viento). Se incluyeron cuatro variables informativas, una redundante y una clasificación binaria, donde la enfermedad se definió como precipitación ≥700 mm, umbral asociado con condiciones de alta humedad (FAO, 2022).

Los datos se dividieron de forma estratificada (train/test) y se aplicó preprocesamiento diferenciado: escalado con StandardScaler para KNN y SVM-RBF, y sin escalado para Random Forest, debido a su independencia de la escala de las variables. Todo el flujo se implementó dentro de pipelines para evitar fugas de datos y garantizar la validez de los resultados (Liakos et al., 2018).

Los parámetros principales fueron:

- KNN: $k = 5$.
- SVM-RBF: kernel radial con pesos balanceados.
- Random Forest: 100 árboles, class_weight='balanced_subsample', random_state=42.

El análisis permitió comparar el rendimiento de tres algoritmos de clasificación con distintos enfoques: KNN, sensible a la escala y el ruido; SVM-RBF, eficaz en separar clases no lineales y minimizar falsos negativos; y Random Forest, robusto y explicativo, aunque con tendencia a priorizar la clase más frecuente.

Resultados y discusión

Resultados Parte 1

El análisis comparativo de los tres modelos de aprendizaje supervisado evidenció diferencias sustanciales en su capacidad predictiva para estimar el rendimiento agrícola a partir de variables climáticas y de manejo. La Tabla 1 resume las métricas obtenidas tras la fase de evaluación, incluyendo el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2), indicadores que permiten valorar tanto la precisión como la capacidad explicativa de cada modelo.

Tabla 1. Resultados comparativos de desempeño de los modelos de regresión.

Modelo	MAE	RMSE	R^2
Regresión Lineal	30.212	40.418	0.649
Gradient Boosting	26.741	36.366	0.716
Random Forest	18.535	24.678	0.869

Los modelos de ensamble (Random Forest y Gradient Boosting) mostraron un ajuste más cercano a la diagonal ideal (y_true = y_pred) en los gráficos de paridad, evidenciando su superior capacidad para reproducir los valores observados de rendimiento. El Random Forest Regressor obtuvo el mejor desempeño ($R^2 = 0.869$; RMSE = 24.678), seguido por Gradient Boosting ($R^2 = 0.716$; RMSE = 36.366) y la Regresión Lineal ($R^2 = 0.649$; RMSE = 40.418). Estos resultados confirman que los métodos de ensamble son más precisos y estables frente a la variabilidad ambiental y de manejo.

La Regresión Lineal, aunque útil como modelo base por su interpretabilidad, presenta limitaciones al asumir relaciones estrictamente proporcionales entre variables. En contraste, los algoritmos de ensamble capturan interacciones no lineales y dependencias jerárquicas entre temperatura, precipitación y uso de pesticidas, reflejando con mayor fidelidad la complejidad de los sistemas agrícolas (Belgiu & Drăguţ, 2016; Nayak et al., 2022).

Estos hallazgos coinciden con lo señalado por Liakos et al. (2018) y Sarker (2021), quienes destacan que los ensambles integran múltiples modelos base para representar mejor la dinámica multivariable del rendimiento agrícola. En síntesis, el Random Forest se perfila como el modelo más equilibrado entre precisión y estabilidad, mientras que Gradient Boosting ofrece un potencial de mejora mediante ajustes de hiperparámetros y Regresión Lineal mantiene valor como referencia interpretativa.

Resultados Parte 2

Las matrices de confusión mostraron que los tres modelos funcionaron razonablemente bien, pero con diferencias claras:

- **KNN**: más errores de todo tipo (FP y FN)
- **SVM-RBF**: menos falsos negativos → mejor recall
- **Random Forest**: menos falsos positivos → mejor precisión

SVM destacó por su capacidad para detectar la mayoría de los casos realmente enfermos, lo que lo hace ideal cuando es crítico no dejar pasar un brote. RF, en cambio, fue más preciso, lo que ayuda cuando las intervenciones tienen un costo alto.

El F1-score fue similar para SVM y RF, pero cada uno priorizó cosas distintas: uno el recall, el otro la precisión. En este tipo de problemas desbalanceados, métricas como PR-AUC y recall suelen ser más útiles que ROC-AUC (Saito & Rehmsmeier, 2015).

En fitosanidad, lo más caro es no detectar una enfermedad. Por eso, aunque el modelo SVM-RBF generó algunas falsas alarmas, su capacidad para detectar correctamente los casos enfermos lo hace el más recomendable. Ahora bien, si actuar ante una alerta tiene un costo alto (como aplicar químicos o frenar exportaciones), quizás convenga usar Random Forest, que lanza menos alarmas falsas. A futuro, ajustar el umbral de decisión según el contexto puede mejorar mucho el equilibrio entre precisión y recall (Chawla et al., 2002).

Resumen de errores en test:

- KNN: FP=103, FN=84
- SVM-RBF: FP=97, FN=61

- RF: FP=77, FN=66

Aunque los resultados son prometedores, hay que tener en cuenta que el dataset es sintético. Habría que validar estos modelos con datos reales de campo y agregar variables más específicas, como humedad en hojas o incidencia de plagas.

También se podrían aplicar técnicas como SMOTE para equilibrar las clases, o usar métodos como SHAP para entender mejor cómo toma decisiones cada modelo. Esto ayudaría mucho en agricultura de precisión, donde no basta con predecir: hay que **explicar** por qué se llega a esas decisiones.

Conclusiones

En este trabajo se entrenaron y compararon modelos de aprendizaje automático supervisado Regresión Lineal, Random Forest y Gradient Boostingaplicados a la predicción del rendimiento agrícola a partir de variables climáticas y de manejo. Los resultados mostraron que los métodos de ensamble, en especial Random Forest, superan a los modelos lineales tradicionales al capturar interacciones no lineales entre temperatura, precipitación y uso de pesticidas, validando su efectividad en la modelación de sistemas agrícolas complejos (Belgiu & Drăguţ, 2016; Nayak et al., 2022).

El proceso metodológico combinó datos reales y sintéticos, permitiendo contrastar el desempeño de los modelos en contextos reales y controlados. En el segundo escenario, centrado en la detección de enfermedades en aguacate Hass, el modelo SVM-RBF destacó por su capacidad para minimizar falsos negativos, priorizando la detección temprana frente a riesgos fitosanitarios.

Referencias.

Allison, P. D. (2001). *Missing data*. Sage Publications.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Food and Agriculture Organization of the United Nations (FAO). (2022). *The state of food and agriculture 2022: Leveraging automation in agriculture for transforming agrifood systems*. FAO. https://www.fao.org

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674. https://doi.org/10.3390/s18082674

Nayak, R., Bhoi, A. K., Mallick, P. K., & Joshi, A. (2022). A comprehensive review on machine learning approaches for crop yield prediction. *Environmental Science and Pollution Research, 29*(24), 36175–36195. https://doi.org/10.1007/s11356-022-19269-8

Patel, J. (2020). *Crop Yield Prediction Dataset*. Kaggle. https://www.kaggle.com/datasets/

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Tripathy, S., Bhoi, A. K., Mallick, P. K., & Zymbler, M. (2020). Advanced machine learning approaches for crop yield prediction and agricultural management. *Cognitive Computation, 12*(6), 1255–1276. https://doi.org/10.1007/s12559-020-09722-7